The Role of Information Retrieval in the Question Answering System IRSAW

Johannes Leveling

Fakultät für Mathematik und Informatik
Intelligent Information- and Communication Systems (IICS)
FernUniversität in Hagen
58095 Hagen, Germany
johannes.leveling@fernuni-hagen.de

Abstract

Information on the internet is a vast resource for question answering. As the amount of available information from web pages increases, novel methods for finding precise answers to user queries and questions must be found. Standard information retrieval methods are efficient, but often fail to provide a user with short, precise answers. A deep linguistic analysis of all information is time consuming, but it offers more advanced means to find answers to a user's question. Shallow natural language processing methods seem to work well on a limited range of questions, but they are not suitable for finding answers to more complex questions.

This paper describes work in progress on the question answering system IRSAW [1](Intelligent Information Retrieval on the Basis of a Semantically Annotated Web), a system that combines information retrieval with a deep linguistic analysis of texts to obtain answers to natural language questions. In IRSAW, different techniques for finding answers lead to different sets of answer candidates, which are then merged to produce a final answer.

The system's architecture and functionality are described before evaluation results of a first prototype are presented.

1 Introduction

The amount of information available on the WWW (world wide web) and information needs of users increase, yet it becomes harder to find relevant answers to questions. Pure information retrieval (IR) approaches fail to provide a user with short, precise answers to information requests. However, IR has managed to scale up with the amount of documents.

Applying natural language processing (NLP) methods to textual information does not scale very well, because a deep linguistic analysis is costly in terms of CPU cycles, but NLP offers a chance to find more precise answers.

Giving short, precise answers to user questions will become more important than returning collections of URLs (Uniform Resource Locators) or whole web pages to read. To this end, combining traditional IR and deep NLP seems to be a promising approach. This paper describes the work in progress on the IRSAW system (Intelligent Information Retrieval on the Basis of a Semantically Annotated Web), a question answering system combining IR approaches with methods for semantic retrieval and logic-based question answering. The project result will be a question answering system capable of answering natural language user questions on the basis of information available on the web.

The IR approach in IRSAW originates from the NLI-Z39.50, a natural language interface to information resources available on the internet [Leveling, 2006a]. It includes features such as blind feedback and query expansion with semantically related search terms and with constituents of compounds (compound nouns are written as one word in German). These methods have been evaluated at the Cross Language Evaluation Forum (CLEF 2003–2006) and achieved a performance of 0.3537 mean average precision (MAP) for the monolingual German domain specific task at CLEF 2006 [Leveling, 2006b].

Methods for question answering (QA) are part of InSicht, a question answering system using a deep linguistic analysis of queries and documents based on a semantic network representation. The InSicht subsystem was first evaluated at the monolingual German QA@CLEF (Question Answering at the Cross Language Evaluation Forum) task in 2004. For the 200 questions of the monolingual German QA@CLEF task in 2005, InSicht found 86 correct answers and 8 inexact answers [Hartrumpf, 2006b]. InSicht is highly oriented towards precision. Only a few inexact and wrong answers were given for the 200 test questions in 2005 and 2006. Furthermore, InSicht has been adapted for a search in web resources [Hartrumpf, 2006a]. Both approaches will be integrated into the IRSAW system.

Current state-of-the art systems often employ IR methods, passage retrieval, or shallow techniques separately or as a combination to pinpoint answers. Neumann et al. [Neumann and Xu, 2003; Neumann and Sacaleanu, 2004] investigate a similar approach, but rely heavily on redundancy of information on the web and on a more expensive preprocessing phase.

Ravichandran and Hovy [2002] suggested and implemented pattern matching based on surface structures (i.e., words) to find answers in the Webclopedia QA system. Since their approach also depends on a specialised and incomplete question/answer typology, it will not be portable

[1]The research presented in this paper was funded by the DFG – Deutsche Forschungsgemeinschaft as part of the project *Intelligent Information Retrieval on the Basis of a Semantically Annotated Web* (IRSAW; GZ: LIS 4 – 554975(2) Hagen, BiB 48 HGfu 02-01).

to open-domain QA.

Ahn, Jijkoun et al. [2006] propose generating several query streams, each returning a set of answers candidates. In contrast to this approach, IRSAW will not merge parallel answer streams, but will favour answer candidates originating from more sophisticated methods (NLP) to those from shallower approaches, such as pattern matching. However, different answer streams will be created in parallel by separate subsystems.

2 Architecture of IRSAW

IRSAW processes user questions in three phases, accessing three kinds of resources: two IR phases in which web search engines and local databases are accessed and a QA phase, in which a semantic network database is accessed. Figure 1 shows the architecture of the IRSAW system. During the first phase, the user question is transformed into an IR query and meta information such as the question type and the expected answer type is determined. The IR query is delivered to web search engines and web portals. Results from the web typically consist of pages with lists of URLs. The web contents (i.e., HTML pages, electronic documents and metadata for audio-visual data) referenced by these URLs are retrieved and converted into text.

In the second phase, the text passages from the web are segmented and indexed in one or more local databases. A local database serves several purposes: First, it is a cache containing texts with answers to previously asked questions and it can be accessed in parallel while the web search is initiated. Second, it is a mediating service with a uniform search interface to heterogeneous services: web systems typically differ in query syntax and in structural elements supported in queries (e.g. support for wildcards, phrases, proximity search). Furthermore, web services may differ in syntax or in the format in which answers are returned (i.e., hierarchical structures or simple lists of URLs). Finally, the local database provides access to units of textual information (text segments) of the same type or length (chapters, paragraphs, sentences, or phrases).

In the third phase of IRSAW, several methods are employed to pinpoint answers. In the InSicht subsystem, a linguistic parser analyses the text segments and semantically annotates them [Hartrumpf, 2003]. The parser returns the representation of the meaning of a text as a semantic network. The semantic representations of questions and texts are compared to intelligently find answers. In addition, a shallower technique creates a different answer stream by applying pattern matching to the answer candidates found in the second phase.

Evaluation of the first prototype of IRSAW is based on data provided for the QA@CLEF task in previous years. The data includes 600 question-answer pairs from 2003–2005 (200 per year) for which missing answers were manually added, and a document collection of 276.581 texts. In the evaluation of the IRSAW prototype, the document collection is a replacement for documents retrieved from the web. The following section describes question processing in IRSAW in more detail. In this paper, the focus is mainly on the IR phases to find answer candidates and pattern matching to find precise answers.

3 Question Processing

3.1 Creating an IR Query

The user poses a natural language question at the client interface, which initiates question processing. For instance an example question might be *"Where was Galileo Galilei born?"*[2]

The natural language question is transformed into an IR query for external web sources (see step 1 in Figure 1). In this process, the set of search terms for the IR query, S, is constructed and optional weights w for search terms are determined. The term weights are only of use for the local database and for search engines supporting a weighted search; they are ignored for the search, otherwise. The set of search terms is constructed from the empty set as follows:

- add all proper nouns and all quoted expressions to the set of search terms to S ($w = 1.0$)
- add all words in upper case (e.g. nouns) to S ($w = 0.9$)
- add all words in lower case (e.g. verbs, adjectives, adverbs) to S ($w = 0.7$)
- add all remaining words (e.g. numeric expressions, etc.) to S ($w = 0.5$)

For the example question, the IR query is *"Galileo_Galilei, born"* with term weights 1.0 and 0.7, respectively.

3.2 Question and Answer Type Classification

Question and answer types are calculated using a Naïve Bayes classifier trained on features representing the first N words of the question. Using $N = 3$ suffices to correctly identify the answer type for 575 of 600 (95.6%) questions in our test corpus. We followed the classification of answers for the QA@CLEF task, defining locations (LOC), persons (PER), organisations (ORG) temporal expressions (TIM), etc. For the example question given above, the answer type location (LOC) is determined.

Question types (see [Helbig, 2006]) include yes-no questions, essay questions, and questions starting with *WH*-words (*why*, *where*, *who*, *when*, ...). The question type will influence the length of the answer and what type of answer is returned.

3.3 Accessing Web Resources

The IR query is sent to external web resources (search engines) which return result pages containing URLs. All web contents referred to by an URL are retrieved and their contents are converted into text. Documents are preprocessed using a sentence and paragraph boundary detection [Grefenstette and Tapanainen, 1994] adapted to German. The resulting texts are then segmented into units and fed to the local database.

3.4 Accessing a Local Database

The IR query is also processed by the local database, which returns a ranked list of text segments. These segments represent answer candidates, i.e. they probably contain an answer to the stated question. If multiple local databases are employed, the local databases and external web resources will be accessed in parallel. Results from the web are added to and indexed in one single local database (using a Round Robin scheme). The question can be processed by

[2] Examples have been translated from German into English.

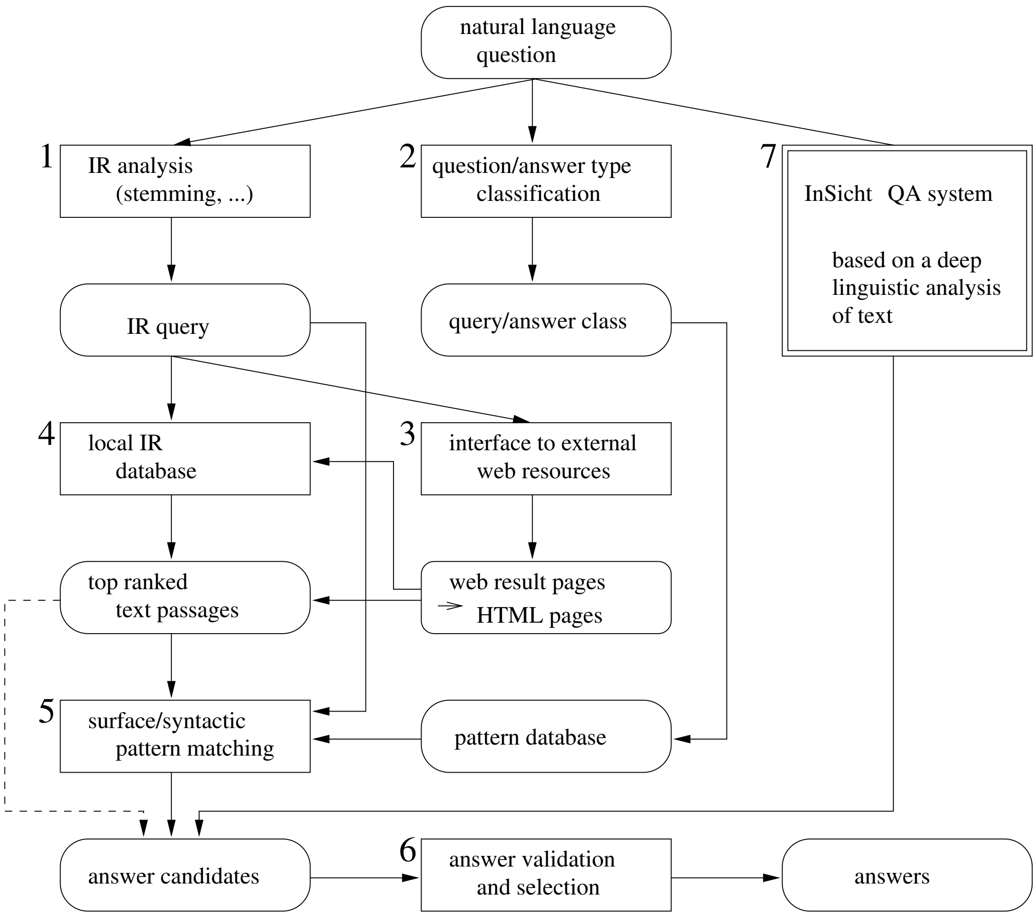

Figure 1: Architecture of the IRSAW system as a combination of IR methods and the QA subsystem InSicht. Arcs indicate data flow, rectangles represent processes, and ovals represent data.

all local databases which are not waiting for web search results and are not busy indexing new results while a web search is in progress.

A promising answer candidate for the example question would be the starting sentence in the Wikipedia article on Galilei: *"Galileo was born in Pisa, in the Tuscany region of Italy on February 15, 1564."* Note that the IR setup described so far already serves as a baseline QA system (indicated by the slashed arc in Figure 1): The top ranked answer candidates obtained from the local database often contain a precise answer if the length of the text unit is adequately chosen.

A pattern database was created using question-answer pairs from previous question answering tasks at CLEF. There were 200 questions for each each of the QA tasks at CLEF in 2003, 2004, and 2005. For these, 732 answers were found – for some questions a text text corpus does not contain a corresponding answer, but for other questions, different correct answers or paraphrases of the same answer are found. Building the pattern database did not involve other manual work such as annotating answers. In the near future, the training data will be largely extended. For a limited range of questions, question-answer pairs can be extracted from available information. One such example is the normalised biographical data about persons (PND – Personennamendatei, see [Hengel and Pfeifer, 2005]) contains information about where and when a person was born and died, and his or her name and professions.

The pattern database contains for each answer type (LOC, PER, ORG, etc.) a number of patterns created from the answer candidates found in the QA@CLEF newspaper corpus. Every answer candidate obtained using the IR query was tokenized and patterns were automatically extracted.

3.5 Pattern Matching

The shallow technique for finding answers in IRSAW is based on pattern matching. A pattern consists of a sequence of variables, symbols, and strings. Variables start with a leading "?". They represent the words at the start and end of a text unit as well as the answer string, matching with zero or more words (tokens). There are two types of symbols: part-of-speech tags (POS tags) from the *Stuttgart-Tübingen Tagset* (STTS) are assigned to words from closed categories (excluding nouns, verbs, adjectives, and adverbs). Other symbols are created from search terms in the IR query that are classified into lower case words (LWORD), upper case words (UWORD), proper nouns (NAME), and numeric expressions (NUM).

The expected answer type for the question is used as a key to find patterns in which the instantiation of the answer variable will be of the expected type. For example, data for the LOC answer type includes all patterns for which the answer variable is a location. Table 1 shows how a pattern for the LOC answer type with a context size of 4 tokens is extracted from the example sentence.

A context window containing a maximum of 5 tokens on each side of the answer variable was used to form the patterns. The pattern matching returns an instantiation of the answer variable (answer string). For example, if the pattern training had been applied to data containing the example answer candidate (*"Galileo was born in Pisa, in the Tuscany region of Italy on February 15, 1564"*), the question (*"Where was Galileo Galilei born?"*), and the answer *"Pisa"*, the corresponding pattern would be *"?words1* NAME ?w0 LWORD appo ?answer+ $comma appo art ?w1 ?words2*"*.

Table 1: Constructing a pattern from the example sentence. The pattern consists of atomic symbols for POS tags, variables corresponding to one or more tokens and special symbols representing words and derived words from the question. The resulting pattern is *"?words1* NAME ?w0 LWORD appo ?answer+ $comma appo art ?w1 ?words2*"*. This pattern has been modified at the start and at the end with a variable matching zero or more tokens (+ denotes a sequence of one or more tokens (words); * a sequence of zero or more tokens).

Text	Tagged Text	LOC pattern
Galileo	NAME	NAME
was	?w0	?w0
born	LWORD	LWORD
in	appo	appo
Pisa	?answer+	?answer+
,	$comma	$comma
in	appo	appo
the	art	art
Tuscany	?w1	?w1
region	?w2	–
of	art	–
Italy	?w3	–
.	$colon	–

All patterns are applied to the top N answer candidates found ($N = 250$) found in the IR phase. Patterns corresponding to matches not containing an instantiation of the answer variable are removed from the list of useful patterns. The remaining patterns are added to the pattern database together with the key (the expected answer type) for lookup.

3.6 Merging, Validating, and Selecting Answers

Answer streams from different sources are merged by preferring answers from InSicht to answers found by the pattern matching. A de-duplication of answer candidates is performed on the answer stream produced by the shallow system. Answers are then ranked by cumulative frequency and the top answer is returned. In the example above, the instantiation of the answer variable *"?answer+"* would be *"Pisa"*.

3.7 The InSicht Subsystem

To produce a second answer stream, IRSAW interfaces to InSicht, a QA system employing a deep linguistic analysis based on a semantic network representation of question and textual information. InSicht has several advantageous characteristics:

- A deep syntactico-semantic analysis for documents and text.
- Independence from other document collection and independence from domains.
- Generation of answers from the semantic network representation of documents, i.e. answers are not extracted from the documents.

InSicht performs best when applied on syntactically correct texts (86 correct and 8 inexact answers were found for 200 questions at QA@CLEF 2005), but it will fail to produce a meaning representation (in this case, a semantic network) for malformed sentences. InSicht's syntactico-semantic parser is able to produce a complete semantic network for about 48.7% and a partial semantic network for for 20.4% of all sentences in the newspaper corpus. Hartrumpf gives an overview over common errors with the WOCADI parser [Hartrumpf, 2005], including the limited robustness of the parser and missing lexicon entries (although the parser relies on a large set of lexicons including full morphological and syntactico-semantic information). For sentences containing grammatical or spelling errors or conflated sentence parts originating from erroneous preprocessing, the parser often fails to produce a semantic network. Thus, InSicht will not to able to find many of those answers appearing in malformed sentences only.

In addition to parser errors and missing lexicon entries, the news articles in the test corpus often contain artefacts from preprocessing as well as metadata such as the date and time of the article, the name of the agency responsible, and the initials of the author conflated into the text. Grammatically, these sentences are not well-formed and thus, any deep linguistic analysis should fail to produce parse results.

However, one should assume that text fragments relevant to an IR query contain to some extent correct answers to questions on the query topic. Therefore, information retrieval methods can be employed a) to interface to IR engines to retrieve textual information for a deeper linguistic analysis, and b) to provide a more robust method to identify answer candidates (because higher ranked answer candidates are more likely to contain an answer).

4 Evaluation Results for the First Prototype

A first evaluation of the IRSAW setup was performed within the question answering task at QA@CLEF 2006. This task consists of finding answers for a set of 200 questions targeting a test corpus of newspaper articles. System answers are assessed manually for correctness. The test corpus contains 276.581 newspaper articles and newswires from the *Frankfurter Rundschau*, *Der Spiegel*, and *Die Schweizerische Depeschenagentur* from the years 1994 and 1995.

A sentence boundary detector and a tokenizer was applied to the test corpus and documents were split into single sentences and indexed in a local database (omitting the phase with web access). Then, question-answer pairs for the QA task in previous years were constructed from the MultiEight corpus [Magnini *et al.*, 2005], augmented manually and used as a training set to build the pattern database.

At QA@CLEF 2006, the pattern matching approach found 17 answers for the 200 test questions, while the method employing deep linguistic processing found correct 61 answers. In total, 64 correct answers were found (one additional answer found by the pattern matching was assessed as inexact). All of the 13 remaining answers found with pattern matching were correct as well. Table 2 shows accuracy and Mean Reciprocal Rank (MRR) for both runs submitted. This was our first approach to combine the deep processing with a shallower method and it leaves many chances for further improvements.

Table 2: Results of the IRSAW QA system for the QA@CLEF task 2006 for 198 assessed questions. Two questions were removed from assessment. (R = right, U = unsupported, I = inexact; A = overall accuracy, MRR = Mean Reciprocal Rank).

QA system	R	U	I	A	MRR
InSicht only	62	4	0	32.28%	32.11
IRSAW + InSicht	65	4	1	33.68%	33.86

The monolingual German QA task in 2006 was more complex in comparison with tasks in previous years, and new types of questions were introduced, e.g. questions including temporal restrictions and list questions. Therefore, a comparison with results from previous years would not be adequate.

The number of correct answers found by IRSAW is expected to increase even more when the shallow method is improved, because the deep linguistic methods were not able to produce answers for some easy questions. However, systems that can be characterised to employ shallower techniques were able to find answers for the same set of topics. The concept of difficult and easy questions is difficult to define, because it depends on the methods employed by a QA system. For easy questions, answers typically are given explicitly in the text (word by word). Answering complex questions will involve paraphrasing and reasoning.

An example shall demonstrate that the overlap in answers produced by the shallow QA subsystem and InSicht is expected to be small, i.e., a substantial performance gain in the combination of deep and shallow processing is likely (see Table 3). For the QA@CLEF question set in 2006, InSicht did not find an answer to a seemingly easy question (topic 0173), but the pattern matching found one. A closer look at the answer snippet reveals that that concept *"satire"* is used metonymically, i.e., it actively participates in or causes an act of failing. This conflicts with the semantic information in the semantic computer lexicon the parser relies on, since a literature genre is not an animated object and therefore can not take the role of an agent. The pattern matching approach simply takes into account the keywords from the question and looks for numeric expressions near them.

In contrast, resolving temporal deictic expressions is beyond the capabilities of a pattern matching approach, but InSicht's inference rules reason at the level of meaning representation. For the first example question (topic 0079), InSicht resolves the temporal deixis *"25 years ago"* to 1970. Viewed from the perspective of the time the article was written (1995), the answer is correct.

The results already show that a combination of processing methods will further improve performance for the IRSAW system. The combination of both traditional IR with deep analysis techniques will provide a highly performing and a more robust system. However, the evaluation task QA@CLEF aims at evaluating a static corpus of newspaper articles. A test targeting web resources has not been performed, yet.

The statement given above is supported from a different view as well. The combination of InSicht with other shallower approach promises a performance boost, because results of the QA@CLEF task show that the performance of a system combining answers (an ideal system) would obtain more correct answers, i.e. there is little overlap between correct answers from systems with different approaches. Magnini et al. [Magnini *et al.*, 2006] give an estimate of a 22% performance increase in accuracy for an ideal system combining results of the monolingual German QA task.

5 Outlook

This paper presented work in progress on IRSAW, a question answering system relying on IR in its initial phases. The evaluation of a first prototype, merging answers from pattern matching based on IR with answers from InSicht was performed with the German question answering task at CLEF 2006 (QA@CLEF). Results for the shallower IR approach (only 17 answers for 200 questions were found) indicate that pattern extraction methods will have to be improved. However, the combination already demonstrates an increase in performance. For the monolingual German QA task at CLEF 2006, IRSAW achieved the second best result of eight monolingual German runs submitted for assessment. Due to increased question complexity, a comparison with results from previous years would not be appropriate.

Future work will include a separate evaluation of the three phases described: retrieval from web resources, retrieval from a local database to find answer candidates, and matching semantic networks to find precise answers.

Major limitations of the shallow approach are the size of the training corpus and the coarse-grained classification of questions and expected answer types. The training corpus will be further extended by large sets of facts extracted from resources available online. Question and answer types will in future be based on semantic properties of the concept representing the answer (i.e., the semantic sort of the concept and its semantic relation to the situation stated in the question). Furthermore, methods involving pattern matching on the surface level and on the syntactic level, used separately and as a combination will be investigated.

References

[Ahn *et al.*, 2006] David Ahn, Valentin Jijkoun, Karin Müller, Maarten de Rijke, Erik Tjong, and Kim Sang. The University of Amsterdam at QA@CLEF 2005. In Carol Peters, Fredric C. Gey, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck, Bernardo Magnini, Henning Müller, and Maarten de Rijke, editors, *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria*, volume 4022 of *Lecture Notes in Computer Science*. Springer, Berlin, 2006.

[Grefenstette and Tapanainen, 1994] Gregory Grefenstette and Pasi Tapanainen. What is a word, what is a sentence? Problems of tokenization. In *3rd International Conference on Computational Lexicography*, pages 79–87, Budapest, 1994.

[Hartrumpf, 2003] Sven Hartrumpf. *Hybrid Disambiguation in Natural Language Analysis*. Der Andere Verlag, Osnabrück, Germany, 2003.

Table 3: Results for two example questions from the QA@CLEF task 2006 (topic 0079 and topic 0173). The first answer was obtained by InSicht, the second answer was produced by pattern matching.

Question	Document sentence	Answer
"In which year did Charles de Gaulle die?"	France's chief of state Jacques Chirac acknowledged the merits of general and statesman Charles de Gaulle, who died 25 years ago.	1970
"In welchem Jahr starb Charles de Gaulle?"	Frankreichs Staatschef Jacques Chirac hat die Verdienste des vor 25 Jahren gestorbenen Generals und Staatsmannes Charles de Gaulle gewürdigt. (SDA.951109.0236)	1970
"In which year was the Russian Revolution?"	The satire inspired by the Russian revolution 1917 lets the dream of liberty and equality fail because of humans.	1917
"In welchem Jahr fand die russische Revolution statt?"	Die von der Russischen Revolution 1917 inspirierte Satire läßt den Traum von Freiheit und Gleichheit an den Menschen scheitern. (FR940612-000533)	1917

[Hartrumpf, 2005] Sven Hartrumpf. Question answering using sentence parsing and semantic network matching. In Carol Peters, Paul Clough, Gareth J. F. Jones, Julio Gonzalo, Michael Kluck, and Bernardo Magnini, editors, *Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004*, volume 3491 of *Lecture Notes in Computer Science (LNCS)*, pages 512–521. Springer, Berlin, 2005.

[Hartrumpf, 2006a] Sven Hartrumpf. Adapting a semantic question answering system to the web. In *Proceedings of the EACL 2006 Workshop on Multilingual Question Answering (MLQA'06)*, pages 61–68, Trento, Italy, April 2006.

[Hartrumpf, 2006b] Sven Hartrumpf. Extending knowledge and deepening linguistic processing for the question answering system InSicht. In Carol Peters, Fredric C. Gey, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck, Bernardo Magnini, Henning Müller, and Maarten de Rijke, editors, *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005*, volume 4022 of *Lecture Notes in Computer Science*. Springer, Berlin, 2006.

[Helbig, 2006] Hermann Helbig. *Knowledge Representation and the Semantics of Natural Language*. Springer, Berlin, 2006.

[Hengel and Pfeifer, 2005] Christel Hengel and Barbara Pfeifer. Kooperation der Personennamendatei (PND) mit Wikipedia. *Dialog mit Bibliotheken*, Jg. 17, H.3:18–24, 2005.

[Leveling, 2006a] Johannes Leveling. *Formale Interpretation von Nutzeranfragen für natürlichsprachliche Interfaces zu Informationsangeboten im Internet*. Dissertation, Fachbereich Informatik, FernUniversität in Hagen, 2006. To appear in: Der andere Verlag, Tönning, Germany, 2006.

[Leveling, 2006b] Johannes Leveling. University of Hagen at CLEF 2006: Reranking documents for the domain-specific task. In Alessandro Nardi, Carol Peters, and José Luis Vicedo, editors, *Results of the CLEF 2006 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2006 Workshop*, Alicante, Spain, 2006.

[Magnini *et al.*, 2005] Bernardo Magnini, Alessandro Vallin, Christelle Ayache, Gregor Erbach, Anselmo Peñas, Maarten de Rijke, Paulo Rocha, Kiril Simov, and Richard Sutcliffe. Overview of the CLEF 2004 multilingual question answering track. In Carol Peters, Paul Clough, Gareth J. F. Jones, Julio Gonzalo, Michael Kluck, and Bernardo Magnini, editors, *Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004*, volume 3491 of *Lecture Notes in Computer Science (LNCS)*. Springer, Berlin, 2005.

[Magnini *et al.*, 2006] Bernardo Magnini, Danilo Giampiccolo, Pamela Forner, Christelle Ayache, Valentin Jijkoun, Petya Osenova, Anselmo Peñas, Paulo Rocha, Bogdan Sacaleanu, and Richard Sutcliffe. Overview of the CLEF 2006 multilingual question answering track. In *Results of the CLEF 2006 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2006 Workshop*, Alicante, Spain, 2006.

[Neumann and Sacaleanu, 2004] Günter Neumann and Bogdan Sacaleanu. A cross-language question/answering-system for German and English. In Carol Peters, Julio Gonzalo, Martin Braschler, and Michael Kluck, editors, *Comparative Evaluation of Multilingual Information Access Systems: 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003*, volume 3237 of *Lecture Notes in Computer Science (LNCS)*, pages 412–424. Springer, Berlin, 2004.

[Neumann and Xu, 2003] Günter Neumann and Feiyu Xu. Mining answers in German web pages. In *Proceedings of the International Conference on Web Intelligence (WI-2003)*, Halifax, Canada, October 2003.

[Ravichandran and Hovy, 2002] Deepak Ravichandran and Eduard Hovy. Learning surface text patterns for a question answering system. In *Proceedings of the 40th ACL conference*, Philadelphia, 2002.